NASA-CR-203449

IN-61-CR
016832

Optical Disk White Paper

Serge Polevitzky[1]

Report RND-90-006, August 1990

Systems Development Department
NAS Systems Division
NASA Ames Research Center
Mail Stop 258-6
Moffett Field, CA 94035-1000

---

[1]Computer Sciences Corporation, NASA Contract NAS 2-12961, Moffett Field, CA 94035-1000

Optical Disk White Paper

SYSTEM DEVELOPMENT DEPARTMENT
Serge Polevitzky

Abstract: Optical disk products are discussed, with emphasis on their applicability to the storage of CFD solution files. WORM (Write Once Read Multiple) and ERASABLE optical disk technologies are compared. Increased conventional disk storage on the workstation, and a Gigabit network are suggested as alternatives to optical disk storage.

I. Introduction

This document (1) provides some information on various optical disk products, (2) explores the value of their possible uses within the NAS community in relation to the petabytes (10^{15}) and exabytes (10^{18}) of solutions anticipated to be produced by NAS[2] .

II. Brief Overview of Optical Disks

Optical disks are similar to conventional disks in that they hold information magnetically. The difference is that conventional disks are written and read with a magnetic head,while optical disks are written with a laser beam of light and read with a photo cell.

---

[2] NAS internal memo, Blaylock to Schairer, dated 9 April 1990, Subject: Background Information Supporting the NAS Augmentation.

Optical disk technology is available in two basic forms: (1) Write Once, Read Multiple (WORM) and (2) "Write Many, Read Many." The latter type is also known by the following aliases: ERASABLE, Thermo-Magneto-Optical [TMO], Magneto-Optical [MO].

WORM is older than ERASABLE, and has been available since approximately 1986. ERASABLE has come to the fore in the last few years, with perhaps the most notable commercial instance being Canon's 256 MByte removable media for the NeXT system. Both WORM and ERASABLE media have a shelf life of approximately 10 years, with data retention of approximately 5 years. (Neither of these limitations would seem to eliminate either WORM or ERASABLE from use at NAS.)

WORM writes data into a spiral track that spins out from the center of the media to the outer edge. ERASABLE format varies, even from one manufacturer. Some types of ERASABLE write data to pie-type segments, others use spiral segments. An ANSI standard for ERASABLE does exist. One can theoretically take a removable ERASABLE disk cartridge written in ANSI format and use it in different vendors' drives, exchanging data between drives of the same or different manufacturers. This is not the case with WORM.

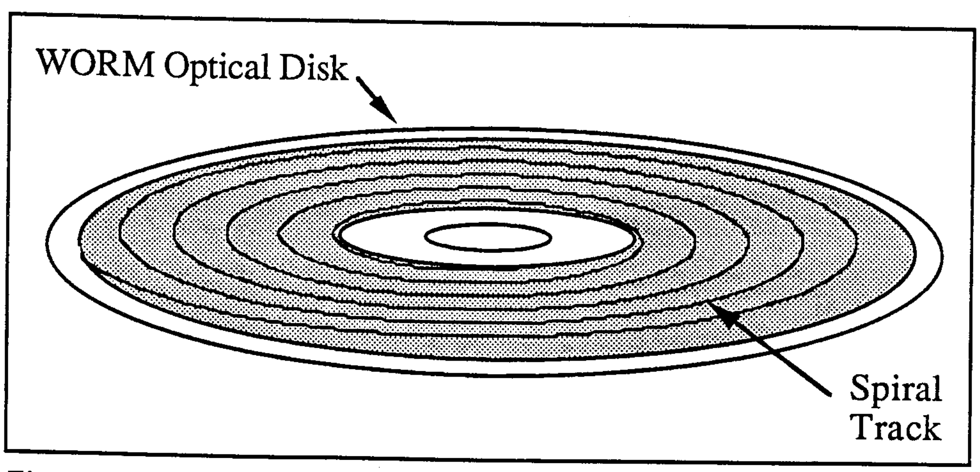

Figure 1. Continuous spiral track on WORM optical disk.

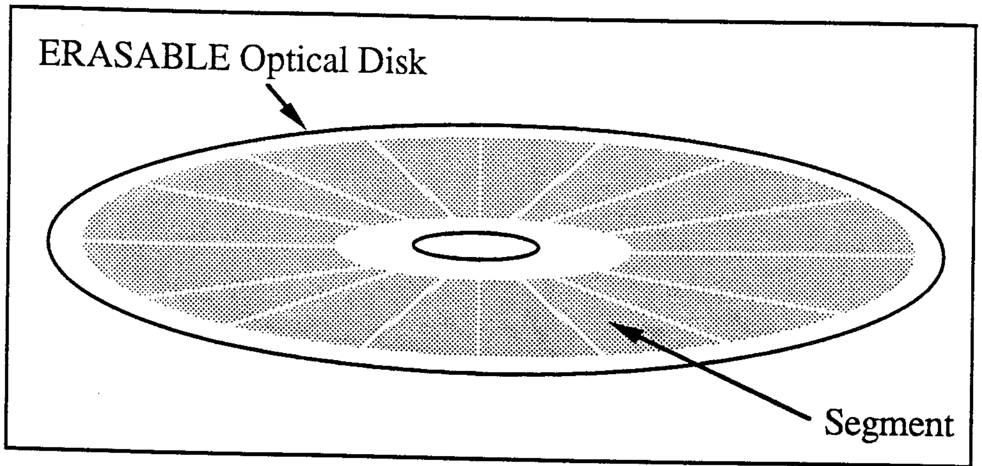

Figure 2. Segments on optical disk.

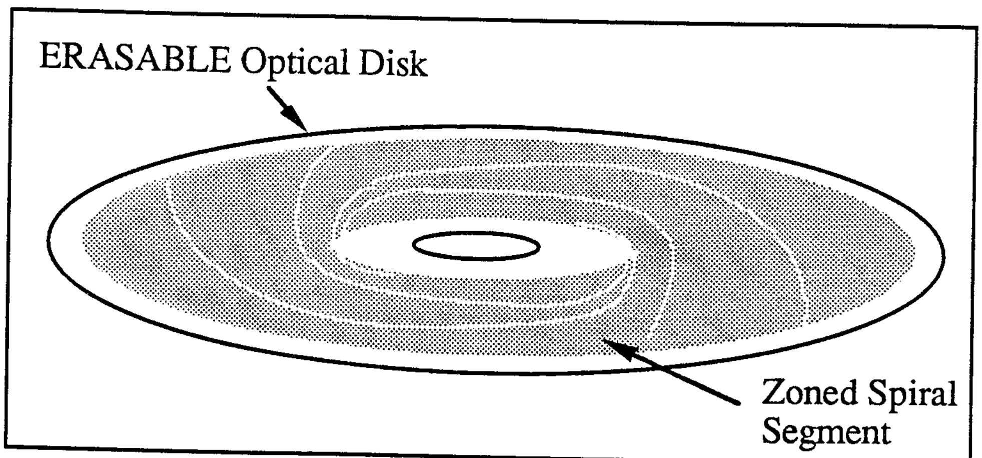

Figure 3. Maxoptix Gigabyte zoned spiral sector erasable optical disk.

III. WORM Optical Disks

WORM is currently thought to be suited only for niche markets (for example, scanned (digitized) images of medical records such as X-rays, entertainment, insurance claims). That is, WORM seems best suited for preserving data that is immutable. The interpretation of your digitized dental X-Rays may change over time, but the digitized X-Rays themselves might be considered frozen data. Compared to conventional disk technology, WORM technology is slow in both seek times and data transfers (see table below), but if the desire is "to write in stone," with no chance to update or modify, then WORM may be suited to your application.

Also, data integrity becomes an issue. If one is saving pixels or bitmaps (for example, a scanned image) a few bits of dropout or a handful of bits picked up out of 13.1 GIGABytes is of little or no concern. If one is saving ASCII or numerical data, however, bits dropped or picked up are critical.

Also, there is a problem in that almost every disk file system is heavily dependent on "adjustments and updates-in-place." Every time Unix reads a file, for example, it does a *write* to update the access date. Thus, reading causes writes. And Unix expects certain important data structures (as in the superblock) always to be at a predetermined sector. The superblock never moves. But WORMs cannot overwrite or write-in-place, they can only update by changing the data and moving it to a never-before-written location. The end result is that WORMs have not faired very well as replacements for conventional magnetic disk. Several commercial file servers (e.g., EPOCH) have switched from WORM to ERASABLE because of the difficulty in emulating conventional magnetic disks with WORMs. A new file system definition is needed; "read only" means read-only.

WORM media defects cannot be detected until the media is written upon. Since the media is Write-Once, the only time that any mechanism can detect a flaw in the media is after the first-and-only write to the media. There is no prequalified or certified WORM media. Some people use the analogy of testing a match or bullet for operational capability: the only way to test a match/bullet is to use it; you cannot pre-test the match or bullet and still use it later. Unix assumes flawless media, so additional code to handle media errors would have to be generated. So there is an implementation challenge!

Flaw detection/mapping during the write process requires that we stop periodically, back up, and verify the data by reading it. The verify operation interrupts the seamless flow of data. Writing without verification does not seem to be worth the risk. Luckily, there is a robust ECC (Error Correcting Code) strategy for correcting data that has become corrupted after being written properly, and many drives use a large (e.g., 256 KByte) on-board buffer. And head crashes are almost eliminated.

The WORM performance data listed below may be considerably higher than the actual performance delivered (see Appendix B).

A large cache or buffer in the host (e.g., several megabytes) to augment the drive's own buffer could reduce the number of backup-and-check-the-data sequences required, but the delivered data rates might be 3/4 to 1/2 those listed below. A prototype system would be mandatory to verify the real delivered data rates.

Table 1.

WORM Performance			
SIZE of MEDIA		5.25"	12"
•	READS	300 KBytes/sec	680 KBytes/sec
•	WRITES	300 KBytes/sec	680 KBytes/sec
•	SEEKS	100 millisec	300 millisec
•	CAPACITY	256 MBytes	6.55 GBytes (single-sided)
•	CAPACITY	550 MBytes	13.1 GBytes (dual-sided)

Observation #1: There are no standards for WORM format, so you and the party with whom you wish to exchange data must choose the same vendor. Hopefully, the vendor chosen will stay in business and not leave you with orphaned media.

Observation #2: With the advent of auto-changers ("jukeboxes"), the ability of one WORM drive to read a disk created by another drive (understood to be from the same vendor) has improved.

Observation #3: Dual-sided drives are becoming available; capacities on the larger drives are, therefore, increasing toward 13.1 GBytes.

Observation #4: To restate an earlier concern: the delivered data rates may be substantially less than those listed in Table 1.

IV. ERASABLE Optical Disks

ERASABLE optical disks are not currently available in formats larger than 5.25". The recording surface for ERASABLE usually consists of segments or spiral segments burned into the media at the factory. Since there are generally two formats for ERASABLE media —ANSI and vendor-specific—a potential for inventory problems exists. A 5.25" Maxoptix disk, for example, could be an ANSI or a Maxoptix-specific format. Since the recording area is pre-formed at the factory, the media cannot be reformatted from an ANSI to a Maxoptix physical layout.

ERASABLE disks require *three* passes under the read-write heads to transcribe data. Because of the relatively long time required to change polarities in the electrical field used to write data, zeros and ones must be written in two separate passes (one full rotational delay between passes, typically 16.6 milliseconds). This is not a problem, only a performance consideration.

Simply put, the first pass is with the electrical field in the up direction (that is, writes all the required zeros in the designated disk sector). A seek back to the start of the desired sector follows, and while the media is being spun around under the read/write heads for a second pass, the electrical field is switched. The second pass with the opposite electrical field (down) enables ones to be written. To insure data integrity, a verification pass is very much recommended (adding a third full rotational delay plus the re-read time). The normal sequence is for the device driver to initialize the microprocessor in the disk controller such that the controller hides all three passes from the host. The utility that writes to the ERASABLE could be more efficient with the verification process by simply buffering larger amounts of data before doing the seek back to the beginning of the last unverified sector to begin the data verify pass. A similar strategy was suggested for WORM devices.

A word of caution about the following table: care needs to be taken since almost every ERASABLE disk vendor quotes transfer rates, and uses the faster performance for the read operation. Again, the likelihood of head crashes is less than with conventional magnetic disks.

Table 2

ERASABLE Performance		
SIZE of MEDIA:	5.25" ANSI	5.25" Vendor-Specific
• READS:	300 KBytes/sec	320 KBytes/sec
• WRITES:	75 KBytes/sec*	90 KBytes/sec*
• SEEKS:	100 millisec†	100 millisec†
• CAPACITY:	550 MBytes	1.0 GBytes
† Exception: Maxoptix Tahiti		
• SEEKS	35 millisec	35 millisec
* Estimated; vendors quote "transfer" times and give read times.		

Observation #1: Even though there are penalties in both capacity and performance for writing ANSI standard format, ANSI offers the flexibility of changing vendors—a very desirable option, especially important if several vendors leave the market.

Observation #2: With the advent of auto-changers ("jukeboxes"), those who wanted more on-line capacity and were willing to pay could be accommodated.

Observation #3: The performance listed is probably realistic.

V. Conventional Magnetic Disks

Conventional magnetic disk technology allows for fairly high-performance transfers (up to 12 MBytes/sec confirmed with IPI-2 drives on an IRIS 320) even if two or more tasks wish to share a disk. Timesharing and multitasking our file systems is easily accommodated by today's conventional magnetic disks.

Table 3

Conventional Magnetic Performance		
SIZE of MEDIA	5.25" SCSI	9" IPI-2
• READS (slow)	1.2 MBytes/sec*	3 MBytes/sec†
• READS (fast)	4.0 MBytes/sec	6 MBytes/sec
• WRITES	Same	Same
• SEEKS	10 millisec	16 millisec
• CAPACITY	1.7 GBytes	2.0+ GBytes

*SCSI supports synchronous (1.25 MBytes/sec) and asynchronous (4.0 MBytes/sec) rates;
†IPI-2 drives are delivered in 3 MBytes/sec and 6 MBytes/sec configurations;
READS and WRITES are accomplished at the same data rates.

SCSI conventional magnetic disks with at least 1.5 GBytes capacity and 1.25 MByte/sec transfer rates are available. 2 MBytes/sec transfer rates have been confirmed with the IO3 board and IRIS 320 striping across two SCSI drives (IRIX 3.3 operating system). This configuration would add 3 GBytes of storage to a workstation.

If two IPI-2 drives were available on the workstation, then over 2.4 GBytes of storage would be added at an honest 4 MBytes/sec transfer rate per drive, 8 MBytes/sec across a striped pair.

Striping across more than two IPI-2 drives may not improve performance because the cpu cannot keep up with the disk bandwidth.

VI. 8-mm Tape

The 8-mm EXAbyte systems provide 5 GBytes of data capacity with 500 KBytes/sec transfer rates. Note that the 8-mm format provides hardware read-after-write, and almost 50% of the bits written to tape are error correcting bits. No seek (or backspace) and re-read what you just wrote is required.

The data rate quoted will, therefore, be very close to the data rate operationally delivered. One drawback is that the media is not available to more than one process at a time. As with optical disks, the user would have to wait until all solution files were written to the media (presumably by a spooler), before the user could access the initial solution file.

A plus for 8-mm is that the EXAbyte 8500 format supports an embedded addressing track that could facilitate a fast search mode (this software would have to be developed).

(Note that DAT (digital audio tape) may be a possibility, too, but that with only a 1.3 GByte capacity and ~200 KByte/sec transfer rates, DAT seems "underpowered.")

VII. Performance Comparison of Optical Disks, Conventional Disks, 8-mm Tape, and Networks.

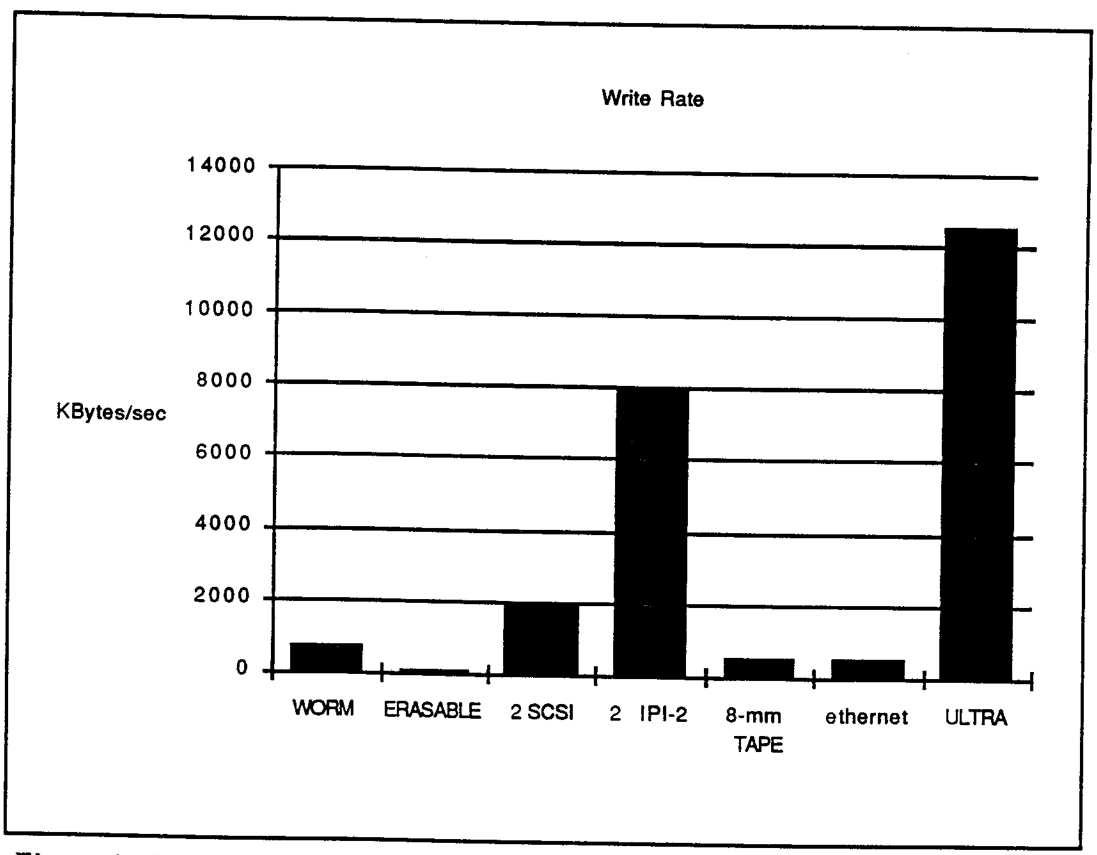

Figure 4. Comparative rates for writing data to media.

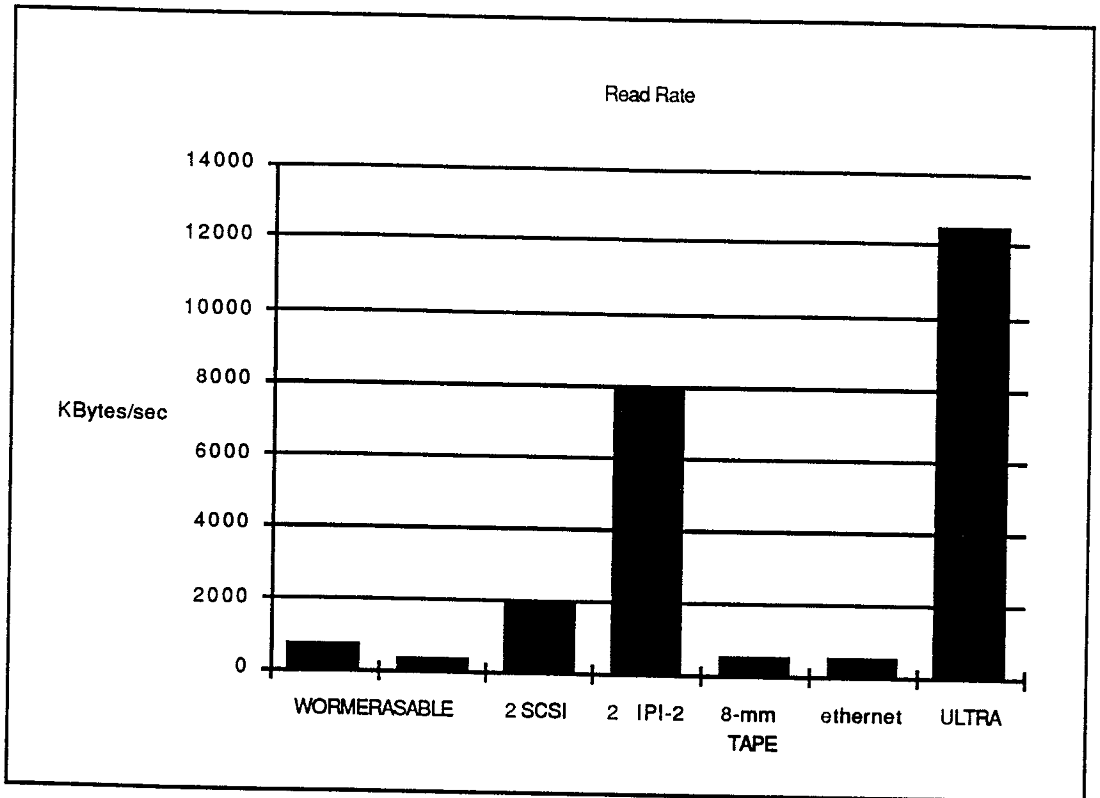

Figure 5. Comparative rates for reading data.

If a scientist uses the High Speed Processor (HSP) to create a solution file, and wants to transfer that file to a workstation for visualization and analysis, a large amount of data needs to be transferred. The following chart compares the time required for writing a 160 MByte file by WORM, ERASABLE, conventional magnetic disks (SCSI and IPI-2), 8-mm tape, ethernet, and ULTRA.

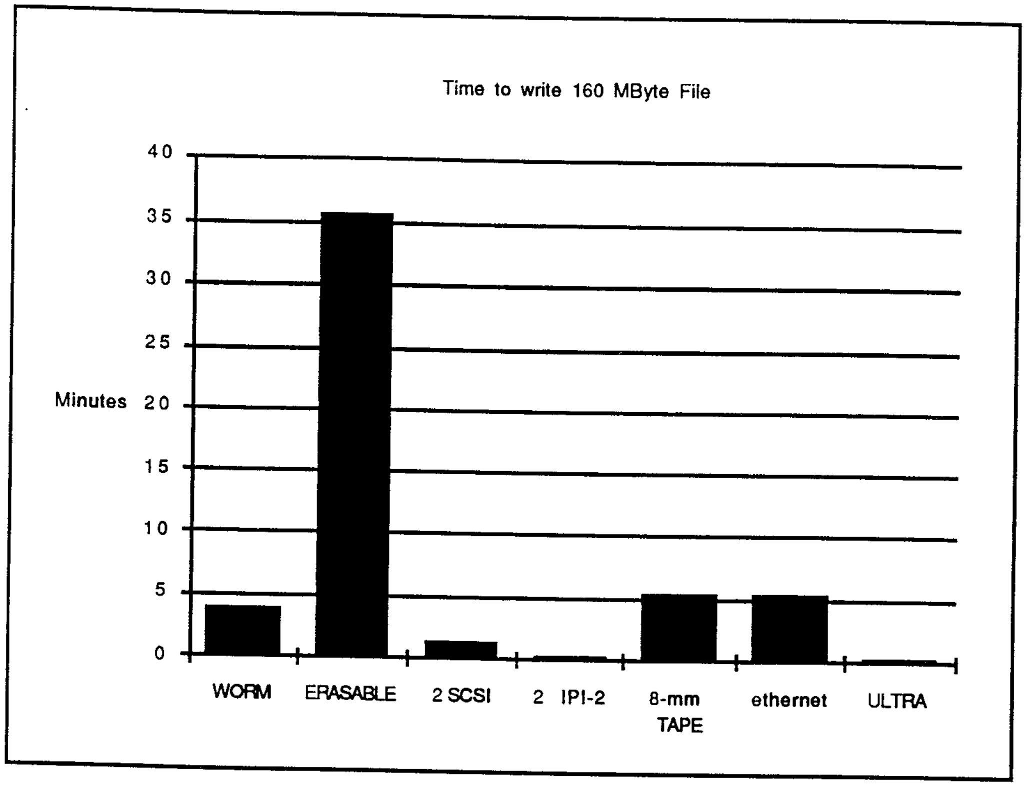

Figure 6. Comparative time needed to write a 160 MByte file.

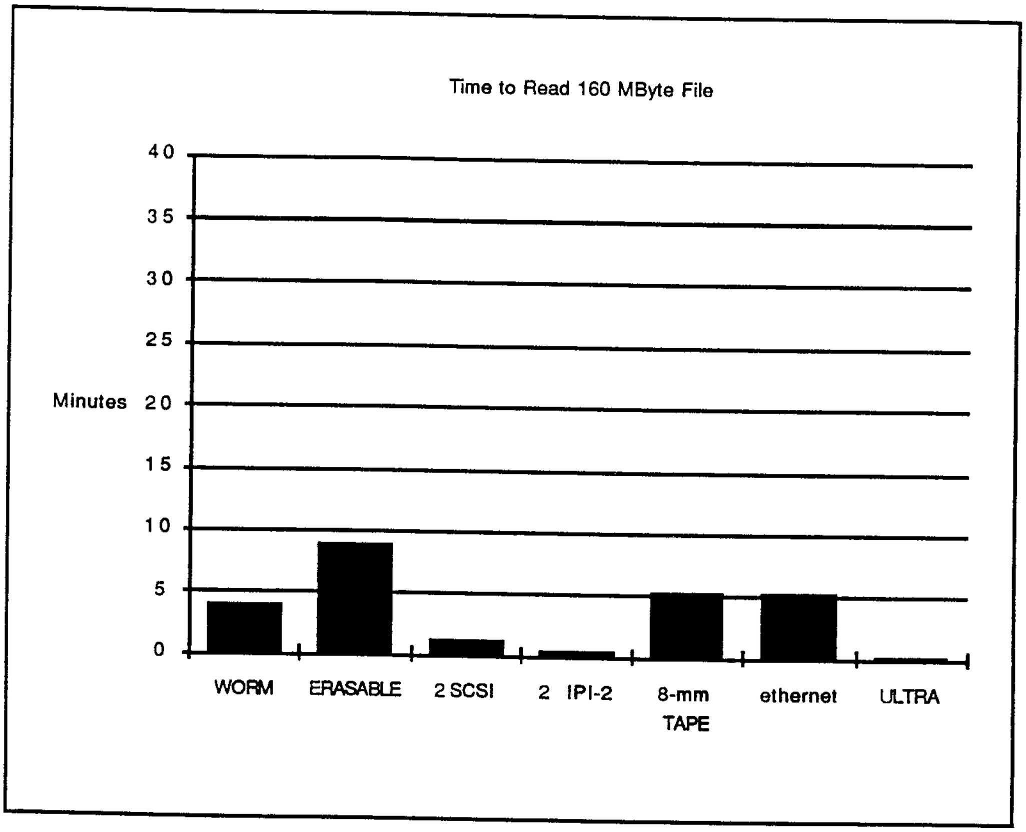

Figure 7 . Comparative time needed to read a 160 MByte file.

VIII. How Optical Disks or 8-mm Tapes Could Be Used at NAS

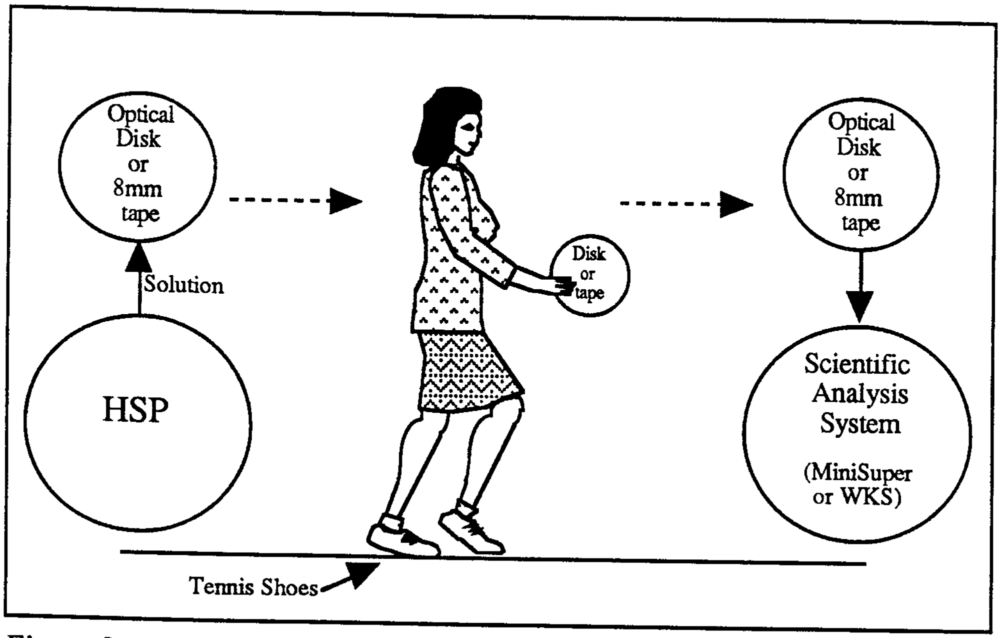

Figure 8. Manual transport of optical disk from high speed processor..

The suggestion has been made that optical disks or 8-mm tapes could be used to store the large solution files created by the HSPs. When a solution file had been written to an optical disk drive attached to the HSP, the scientist would pick up the disk or tape and carry it back to his or her workstation (or minisuper), where the solution would be visualized.

The advantages suggested by this method might be:

- The file could be removed from the Cray.
- Stress on the network would be relieved.

The disadvantages are:

- The process would be slower than a good network.
- The scientists would wind up with drawers full of disks, all vulnerable to the Coke-spill syndrome (but networks also can trash file).
- The scientists would have to take the time to master a new technology.
- Exchanging data with off-site researchers would be hampered by incompatible technologies.
- Optical disks and 8-mm tapes are not yet as cheap or fast as conventional disks.

IX. Network Alternative

An alternative to adding optical disk drives to the HSPs and the workstations is to add more muscle to the network, and add striped conventional drives to the workstations.

160 MByte files transferred across the network every time the Crays can produce a solution file might well flood the existing network. Here are some estimates as to network loading:

Assume each Cray produces a 160 MByte file every 10 minutes, or 266 KBytes per second per HSP. Two HSPs means approximately 533 KBytes (0.533 MBytes) of solutions every second. This means for each 3600 seconds (or one hour) the two HSPs produce approximately 1.918 GBytes. That equals 46 GBytes/day. New Massively Parallel Processors will be added to NAS and will add to the solution file production.

Assume the solution files get created, and then each Cray tries to deliver the file as fast as possible to the workstation. Each Cray might grab for several successive and large blocks of the network bandwidth, and then go quiescent, until the next solution file were finished.

Solution files will be transmitted to the workstations (or to the SAS, the proposed Scientific Analysis System which will aid in the understanding of CFD solutions), and they will also have to be transmitted to the node that will store them (Cray to MSS-2 transfer, for example). So each solution file will probably be transferred at least twice—once to the workstation, and once to MSS-2. This suggests a High Speed Network connecting just the HSPs, MSS, and SAS, for just the "high traffic, fast data" producers, and a separate Medium Speed LAN for users; that is, a hierarchy of networks.

If a Gigabit network were to be available, the operational bandwidth would be approximately a factor of 10 to 20 over ethernet. ULTRA, for example, states that it will deliver an operational data rate of at least 50% of the raw bandwidth via its TP4 protocol engine(s). This would yield an available, operational 50 MBytes/sec from the Crays' HSX channels, rated at 100 MBytes/sec. If we accept ULTRA's no-more-than-50% overhead statement at face value, but then discount the ULTRA claim by 75%, we can still expect a 12.5 MBytes/sec transmission rate.

Our Gigabit network would have a bandwidth budget of 12.5 MBytes/sec. This would mean the ability to move more than a terabyte of data each 24 hours at 12.5 MBytes/sec. Data would be stored in MSS 2.

The Crays are always producing their solution files, always producing output at 0.533 MB/s. Until we have an SAS, we should double this rate to allow for movement of the solution files to MSS-2 from the workstations: $2 * 0.533$ MB/s = approximately 1.06 MB/s. Doubling the 1.06 MB/sec for headroom (bandwidth budgeting) would yield ~ 2.2 MBytes/sec. 12.5 MBytes/sec — 2.2 MBytes/sec = 10.3 MBytes/sec. Our Gigabit network would have a balance of ~ 10.3 MBytes/sec for other traffic. This amounts to over 864 GBytes available to other processes per 24 hours.

If a Gigabit network (or some other solution) cannot supply the necessary bandwidth for transmitting the solution files from the HSPs, then a manual transfer of data should be considered, or several subnets of Gigabit networks.

X. Conclusions

Table 4

Optical and Network Performance Time to Write a 160 MByte Solution File		
MEDIA	Data Rate	Time
• WORM	680 KBytes/sec	4 minutes (235 secs)
• ERASABLE	75 KBytes/sec	35 minutes (2133 secs)
• EXAbyte 8500	500 KBytes/sec	5 minutes 20 secs (320 secs)
• ethernet	.5 MByte/sec	5 minutes 20 seconds (320 secs)
• ULTRA	2.2 MByte/sec	1 minute 12 seconds (72 secs)

The WORM and ERASABLE solution seems to suffer from marginal data transfer rates, a dearth of *de facto* or real standards or strategies aimed at the higher-performance end of the market. There are no projects to adopt VME or IPI-2 or IPI-3 to either WORM or ERASABLE. The 1.25 MBytes/sec Asynchronous SCSI seems to be satisfactory to the optical vendors. In fact, rather than faster or larger size and larger capacity than current 5.25" ERASABLE drives, the 3.5" drives seem to have the vendor's attention.

WORM may be able to deliver the performance sought. Again, the delivered performance for either reading or writing may be significantly lower than the performance data quoted above.

Once again, before any commitment to a new media or technology is made, NAS should prototype several of the more-promising hardware-software offerings before creating a library of Terabytes of information.

In conclusion:

#1 The Gigabit network really is the solution. Or at least it should be tried first.

#2 Add local disk storage capacity to workstations whether or not the network can handle the bandwidth. Stripe across drives. Upgrade disk controllers where practical. Consider IPI over SCSI where cost will permit.

#3 If the Gigabit network cannot deliver the required bandwidth for the solution-file traffic, then place 8-mm tape drives on the workstations and install several SUNs or SGIs with FEI-3 interfaces at the end of CRAY low speed channels (LSPs). Develop spoolers to unload the solution files to the 8-mm tapes, advising users when to start walking so they can pick up their solution tapes/files and walk back to their workstations.

#4 If 8-mm tape, WORM, or ERASABLE is selected, a prototype system (returnable if possible) should be tried before a large number of drives are acquired. We would need to verify operational data transfer rates, device driver robustness, and user acceptance of this strategy.

#5 MSS 2 would store the data.

Appendix A. Vendor Data

Table 5: Optical Drive Systems available for Iris (23 October 1989)

WORM Subsystems

Vendor	Model	Interface	Capacity unformatted (Mbytes)	Avg. Seek Time (msec)	R/W Transfer Rate (Kbytes/sec)	Price Unit/Media	Price/Mbyte
ATG Gigadisc	GD1002	VME	2000	115	470/200	$12,300.00/ $550.00	$6.43
ATG Gigadisc	GD6000	VME	6400	90	990/990	$18,600.00/ $820.00	$3.03
Introl Corporation	Sterling 654W	VME	654	60	100	$8720.00/ $156.00	$13.33
Q-Systems SGI-WOSD	OFS/	SCSI	3200		150/80	$28,500.00/ $360.00	$9.02
Aquidneck	OAS	VME	3200		100	$36,932.00/ $337.00	$11.64

Erasable Optical Subsystems

Vendor	Model	Interface	Capacity unformatted (Mbytes)	Avg. Seek Time (msec)	R/W Transfer Rate (Kbytes/sec)	Price Unit/Media	Price/Mbyte
Introl Corporation	Sterling 650E	VME	650	95	50	$10,520.00/ $236.00	$16.18
Genesis Imaging	S-501	SCSI	650	95	690	$5,995.00/ $230.00	$9.57
Q-Systems	OFS/ SGI-MOSD	SCSI	650	90	300/160	$7500.00/ $250.00	$11.92

Multidisc Subsystems

Vendor	Model	Interface	Capacity unformatted (Gigabytes)	Avg Access Time (msec)	R/W Transfer Rate (Kbytes/sec)	Price Unit/Media	Price/Mbyte
Epoch Systems	Epoch-1 model 33	NFS	31.3 (up to 150)		300	$162,700.00	$5.20
ATG Gigadisc	GD6000	VME	704 (up to 912.4)	10	990/990	$315,200.00	$0.45
Q-Systems (erasable)	OFS/ SGI-MOJB	SCSI	20	3	300/160	$52,500.00	$2.62
Q-Systems	OFS/ SGI-WOJB	SCSI	160	3	150/80	$159,000.00	$0.99
Aquidneck Systems International	OAS	VME	164 (up to 1000)		100	$140,500.00	$0.86

Table 6: Available for Sun; drivers could be ported

(23 October 1989)

WORM Subsystems

Vendor	Model	Interface	Capacity unformatted (Mbytes)	Avg. Seek Time	R/W Transfer Rate (msec)	Price Unit/Media (Kbytes/sec)	Price/Mbyte
Delta Microsystems*	SS-622W	SCSI	622	60	269/110	$6000.00/ $150.00	$9.89
Delta Microsystems*	SS-2000WA	SCSI	2000	150	220/150	$18,750.00/ $550.00	$9.65
General Microsystems	OL/D440	VME	2400	150	334	$21,150.00/ $599.00	$9.06

Erasable Optical Subsystems

Vendor	Model	Interface	Capacity unformatted (Mbytes)	Avg. Seek Time (msec)	R/W Transfer Rate (Kbytes/sec)	Price Unit/Media	Price/Mbyte
Pinnacle Micro	REO-650	SCSI	650	65	650	$4,790.00/ $250.00	$7.75
RELAX Technology*	Erasable Optical Plus	SCSI	570 (formatted)	67	693	$3095.00/ $225.00	$5.53
Maxtor*	Tahiti I	next yr	1000	35	800/400		
SUMMUS Computer Systems	LightDisk 650	SCSI	650	90	620	$5600.00/ $250.00	$9.00
Alphatronix*	Inspire	SCSI	650	83	325	$6995.00/ $250.00	$11.15

Multidisc Subsystems

Vendor	Model	Interface	Capacity unformatted (Gigabytes)	Avg Access Time	R/W Transfer Rate (msec)	Price Unit/Media (Kbytes/sec)	Price/Mbyte
Pinnacle Micro (erasable)	REO-16000	SCSI	16		650	$63988.00	$3.99
Alphatronix*	Inspire	SCSI	30		325	$74,900.00	$2.49
Delta Microsystems*		SCSI	up to 200			$250,000.00	$1.25

* Vendors are willing to port their device driver to the Iris or are willing to give us source or detailed specifications .

Appendix B
Sustained Data Rates

A simple yardstick for estimating a disk's performance is to look at the number of sectors per track and the number of surfaces in the Head Disk Assembly, or HDA. The disk cannot read or write more sectors than pass under the read-write head(s), so if we know the number of sectors per revolution and know the spin rate, we can find an upper bound as to the number of sectors (hence, bytes) that can be read or written per revolution.

Some disks vary the number of sectors per track, packing more tracks at the edge than toward the center. So it is often possible to read and write more data per revolution at the edge than at the center. Hence, data transfer rates quoted for the "edge" can be higher than for middle or the inner tracks.

Below is an example of how the manufacturer's quoted sustained data transfer rate may be higher than the "real" sustained data rate, and the manner in which both manufacturer's and "real" were derived:

130mm (5.25") ISO/ANSI Standard MO (ERASABLE) disk —

Manufacturer's Quoted User Data Transfer Rate:
680 KBytes/sec (1024 byte/sector) — 17 sectors/tracks
620 KBytes/sec (512 byte/sector) — 31 sectors/tracks
[Source, Tecmar LaserVault™ brochure dated 5/90]

How did Tecmar come up with their stated Data Transfer Rate ? Well, a single-headed disk has 17 sectors / track at 1024 Bytes/sector. Rotational speed equals 2400 RPM (= 40 revolutions/second).

40 revolutions/sec * 17 sectors/revolution * 1K bytes/sector = 680KBytes/second

Or, for 512 byte sectors (31 sectors/track):

40 revolutions/sec * 31 sectors/revolution * 0.5K bytes/sector = 620 KBytes/second

So we can backtrack to derive the manufacturer's quoted sustained data rates. But are these what we should expect for sustained data rates for the file sizes we will be creating?

Consider that both the 620 or 680 KByte data transfer rates can be sustained for only 1 full revolution (or 1/40th of a second = 25 milliseconds). That is, either 31 or 17 sectors are transferred every rotation—every 25 milliseconds. The read-write head must then seek to the next track (hopefully, a short distance), and then wait for the rotational latency until the desired next-beginning sector appears under the heads.

A seek latency of 25 milliseconds and half the 25 milliseconds rotational latency is added to each full sector transfer. This translates into 25 + 12.5, or 37.5 milliseconds of overhead for every 25 milliseconds of "real" reading or writing. Real, sustained data throughput for a 160 MegaByte file becomes something more on the order of (1024 bytes/sector):

17 sectors transferred every 25 milliseconds (real)+ 37.5 milliseconds (overhead, to set up for the next transfer), or 17 sectors * 1 KB/sector for every 25.0+37.5 milliseconds, or 17KB every 62.5ms, or multiplying by 16 to get seconds = 272KB/sec.

Remember, the vendor qouted 680 KBytes/sec for this drive.